DOCUMENT RESUME

ED 380 495 TM 022 858

AUTHOR Zwick, Rebecca; And Others
TITLE Effect of Rasch Calibration on Ability and DIF Estimation in Computer-Adaptive Tests. Research Report RR-94-32.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE May 94
NOTE 40p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability; *Adaptive Testing; *Computer Assisted Testing; Computer Simulation; Educational Assessment; Estimation (Mathematics); Identification; *Item Bias; Item Response Theory; Test Items; *True Scores
IDENTIFIERS *Calibration; Mantel Haenszel Procedure; *Rasch Model; Standardization; Three Parameter Model

ABSTRACT

A previous simulation study of methods for assessing item functioning (DIF) in computer-adaptive tests (CATs) showed that modified versions of the Mantel-Haenszel and standardization methods work well with CAT data. In that study, data were generated using the three-parameter logistic (3PL) model, and this same model was assumed in obtaining item parameter estimates. In the current study, 3PL item response data are used, but the Rasch model is assumed in obtaining item parameter estimates. Rasch-based expected true scores are obtained for each examinee. As in the previous study, DIF statistics are highly correlated with the generating DIF, and the means and standard deviations across items are close to their nominal values. There is, however, a tendency for DIF statistics to be slightly smaller in magnitude than in the 3PL analysis, resulting in a lower probability of detecting items with extreme DIF. Expected true scores from the Rasch-based CAT tend to be biased downward, particularly for lower ability examinees. Unlike Rasch CAT scores, Rasch expected true scores based on nonadaptive administration of all pool items behave well, as do nonadaptive and CAT-based expected true scores obtained using the 3PL model. Eight tables and two figures present analysis information. (Contains 16 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-94-32

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. COLEY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

EFFECT OF RASCH CALIBRATION ON ABILITY AND DIF ESTIMATION IN COMPUTER-ADAPTIVE TESTS

Rebecca Zwick
Dorothy T. Thayer
Marilyn Wingersky

ED 380 495

Educational Testing Service
Princeton, New Jersey
May 1994

TM022858

Effect of Rasch Calibration on Ability and DIF Estimation in Computer-Adaptive Tests

Rebecca Zwick, Dorothy T. Thayer, and Marilyn Wingersky

Educational Testing Service

April 21, 1994

We thank Charles Lewis, Gary Marco, Robert Mislevy, Nancy Petersen, and Denny Way for helpful discussions and reviews. We appreciate the sponsorship of the Program Research Planning Committee at ETS.

Copyright © 1994. Educational Testing Service. All rights reserved.

Abstract

A simulation study of methods of assessing differential item functioning (DIF) in computer-adaptive tests (CATs) was conducted by Zwick, Thayer and Wingersky (in press; 1993). Results showed that modified versions of the Mantel-Haenszel and standardization methods work well with CAT data. In that study, data were generated using the three-parameter logistic (3PL) model and this same model was assumed in obtaining item parameter estimates. In the current study, 3PL item response data were used, but the Rasch model was assumed in obtaining item parameter estimates, which, in turn, determined the information table to be used in the item selection algorithm. New Rasch-based expected true scores were obtained for each examinee, based on responses to the CAT items. As in the previous study, the DIF statistics were highly correlated with the generating DIF, and the means and standard deviations of these statistics across items were close to their nominal values. There was, however, a tendency for DIF statistics to be slightly smaller in magnitude than in the 3PL analysis, resulting in a lower probability of detecting items with extreme DIF. This reduced sensitivity appeared to be related to a degradation in the accuracy of matching. Expected true scores from the Rasch-based CAT tended to be biased downward, particularly for lower-ability examinees. Unlike the Rasch CAT scores, Rasch expected true scores based on nonadaptive administration of all pool items behaved quite well, as did the nonadaptive and CAT-based expected true scores obtained using the 3PL model.

1. Overview

Zwick, Thayer, and Wingersky (in press; 1993; henceforth referred to as ZTW) conducted an extensive study of DIF methods for CATs. Simulated data were used to investigate the performance of modified versions of the Mantel-Haenszel (MH; 1959) approach of Holland and Thayer (1988) and the standardization method of Dorans and Kulick (1986). Each simulated "examinee" received 25 items out of a 75-item pool. For DIF analysis, examinees were matched on expected true scores based on their CAT responses and on estimated item parameters. Both DIF methods performed well. The CAT-based DIF statistics were highly correlated with DIF statistics based on nonadaptive administration of all 75 pool items and with the true magnitudes of DIF in the simulation. In addition, the across-item means and variances of the DIF statistics were close to their nominal values.

In ZTW, data were generated using the 3PL model; this same model was assumed in obtaining item parameter estimates. In at least one nationally administered CAT, however, item parameter estimates based on the Rasch model are being used. Application of the Rasch model to multiple-choice items is controversial because the model incorporates the assumption that items cannot be answered correctly through guessing. The present study investigated the effect on DIF results of using the Rasch model to obtain item parameter estimates when the data have been generated with the 3PL model. The study also allowed the investigation of a more fundamental issue: the effect of Rasch calibration on the quality of ability estimation.

As a first step, item response data from ZTW, which included responses to all items in the CAT pool, were used to recalibrate the items, assuming the Rasch model. A new item information table was constructed using the Rasch-based estimates of item difficulty, and the CAT was then "readministered" to each examinee. New ability estimates and expected true scores were obtained. Finally, examinees were matched on the new scores, and DIF statistics were computed. The results of the study consist of two main components: a comparison of the Rasch-based expected true scores to the 3PL scores and to the corresponding true values, and a comparison of the DIF results to those previously obtained and to the true DIF used in data generation.

The sections that follow give, in turn, a description of the simulation conditions, a discussion of the procedures used for ability and DIF estimation, a presentation of the results of each of these two components of the study, and a summary.

2. Simulation Conditions

DIF analyses are typically based on two groups--the group of primary interest, or *focal* group, and the group to which it is compared, or *reference* group. In this study, three different focal groups were compared, in turn, to the reference group. The ability distributions for all groups were normal, with a standard deviation of 1. The reference group distribution and one of the focal group distributions had means of 0. The other two focal distributions had means of -1 and +.5, respectively. The differences between reference and focal group means were chosen to be representative of group differences encountered in ETS DIF analyses.

The three focal group distributions, along with the two item pools (see section 2.1) and two sample size conditions (see section 2.2) resulted in a total of 12 simulation conditions, summarized in Table 1.

Insert Table 1 about here.

2.1. CAT Item Pools

In the present study, responses to two different pools of 75 items were simulated. In Pool 1, the items had no DIF; in Pool 2, the items had DIF that was uncorrelated with item difficulty. (A third pool used in ZTW, in which the items had DIF that was correlated with item difficulty, was not included in the present study.)

The factors that were varied across the items were the item discrimination parameters (a), difficulty parameters (b), and, in the case of Pool 2, DIF parameters (d). The DIF parameter for item j was defined as $d_j = b_{jR} - b_{jF}$, where $b_{jR} = b_j$ is the reference group difficulty and b_{jF} is the

focal group difficulty. Therefore, a value of *d* greater than zero implied that an item was easier for the focal group than for the reference group, whereas *d* less than zero implied that the item was harder for the focal group.

Estimates of item parameters and DIF from actual admissions test data were used in determining the values of *a*, *b*, *c*, and *d* for the simulation. Values of ln (*a*), *b*, and *d* were sampled from a multivariate normal distribution. (See ZTW for details.) To simplify the simulation, the guessing parameter, c_j, was set equal to .15 for all items. Item responses were generated with the three-parameter logistic (3PL) item response function (Birnbaum, 1968), using the true item and ability parameters. The included values of *a*, *b*, and *d* for the items were as follows:

a: .74, 1,

b: -1.95, -1.3, -.65, 0, .65, 1.3, 1.95, and

d: 0 in Pool 1; -.70, -.35, 0, .35, .70 in Pool 2.

The CAT simulation was designed as a simplified version of actual CATs being developed at ETS. The CAT algorithm, originally developed for the ZTW study, selects as the next item to be administered the most informative item at the maximum likelihood estimate (MLE) of ability computed from the items already administered. (Unlike most operational CAT algorithms at ETS, the simulation algorithm does not take into consideration such factors as item format, content, and exposure rate when selecting items for administration.) The item parameter estimates used for computing item information and ability estimates were obtained through an analog to a paper-and-pencil test administration. A sample of 2,000 reference group examinees were "administered" all 75 items, and the LOGIST program (Wingersky, 1983; Wingersky, Patrick, & Lord, 1988) was used to estimate the *b* parameters. The distribution of examinee ability estimates was constrained to have a mean of zero and a standard deviation of one. DIF was induced in this study by manipulating the *focal group* difficulties only; therefore, the data from the calibration sample do not contain any DIF.

Because 2,000 is a typical sample size for such calibrations, this approach allowed for the incorporation of a realistic amount of estimation error. The true *a*, *b*, and *d* parameters, along with

the estimated *b* parameters, are given in Table 2. (The value of *ad* for each item is also tabled for reasons explained in section 6.1.) As is typical when the Rasch model is applied to 3PL data, the difficulties of the items are systematically underestimated as a result of the assumption that items cannot be answered correctly through guessing.

Insert Table 2 about here.

2.2. Sample Size Conditions

Two sample size conditions were of interest: $n_R = 500$, $n_F = 500$ and $n_R = 900$, $n_F = 100$, where n_R and n_F are the sample sizes for the reference and focal groups, respectively. These sample size conditions, which were also used in ZTW, were chosen to be similar to those that occur in DIF analyses conducted by ETS testing programs. Further discussion of sample size issues is given in section 4.3.

3. Procedures for Estimating Ability and Matching Examinees for DIF Analysis

In DIF analysis, examinees are first grouped on the basis of a matching variable that is intended to be a measure of ability in the area of interest. In many DIF applications, the matching variable is the total score on the test in which the studied item is embedded. In ZTW and in this study, the matching variable for the DIF analysis of the CAT-administered items was obtained by (1) getting the examinee's MLE of ability, based on the responses to the 25 CAT items and (2) using this MLE, along with the LOGIST estimates of the item parameters, to compute an expected true score for the entire item pool by summing the 75 values of the estimated item response functions. That is, the matching variable is

$$CAT\ expected\ true\ score = \sum_{j=1}^{75} \hat{P}_j(\hat{\theta}_{CAT}), \qquad (1)$$

where $\hat{P}_j(\hat{\theta}_{CAT})$ is an estimate of the item response function for the *j*th item and $\hat{\theta}_{CAT}$ is the MLE of ability based on the CAT items. The difference across studies in the matching variable lies in the definition of $\hat{P}_j(\hat{\theta}_{CAT})$. In ZTW, $\hat{P}_j(\hat{\theta}_{CAT})$ was an estimate of the 3PL function based on 3PL estimates of ability and item parameters, whereas in the present study, $\hat{P}_j(\hat{\theta}_{CAT})$ was an estimate of the Rasch item response function based on Rasch estimates of ability and item parameters. Examinees whose values on the matching variable fell in the same one-unit intervals were considered to be matched. In the 3PL study, expected true scores ranged from $\sum_{i=1}^{75} \hat{c}_i$ (which was equal to 10.7) to 75; in the Rasch study, they ranged from 0 to 75.

4. DIF Methods

This section describes the two DIF methods and the procedure used to estimate the behavior of DIF statistics under various sample size conditions. The score on the studied item, group membership, and the value of the matching variable for each examinee define a 2 x 2 x *K* cross-classification of examinee data, where *K* is the number of levels of the matching variable. This contingency table serves as the basis of the MH and standardization DIF methods.

Assume that there are T_k examinees at the *k*th level of the matching variable. Of these, n_{Rk} are in the reference group and n_{Fk} are in the focal group. Of the n_{Rk} reference group members, A_k answered the studied item correctly while B_k did not. Similarly C_k of the n_{Fk} matched focal group members answered the studied item correctly, whereas D_k did not.

4.1. Mantel-Haenszel DIF Analysis

The MH measure of DIF is

$$MH\ D\text{-}DIF = -2.35 \ln(\hat{\alpha}_{MH}) \qquad (2)$$

where $\hat{\alpha}_{MH}$ is the Mantel-Haenszel conditional odds-ratio estimator given by

$$\hat{\alpha}_{MH} = \frac{\sum_k A_k D_k / T_k}{\sum_k B_k C_k / T_k} \tag{3}$$

In equation 2, the transformation of $\hat{\alpha}_{MH}$ places *MH D-DIF* on the ETS delta scale of item difficulty (Holland & Thayer, 1985). The effect of the minus sign is to make *MH D-DIF* negative when the item is more difficult for members of the focal group than it is for comparable members of the reference group. An estimated standard error for *MH D-DIF*, based on work by Robins, Breslow and Greenland (1986) and Phillips and Holland (1987), is given in Holland and Thayer (1988). It is

$$SE(MH\ D-DIF) = 2.35\sqrt{Var(\ln(\hat{\alpha}_{MH}))} \tag{4}$$

where $Var(\ln(\hat{\alpha}_{MH}))$ is estimated by

$$\frac{\sum_k U_k V_k / T_k^2}{2\,(\sum_k A_k D_k / T_k)^2}, \tag{5}$$

where $U_k = (A_k\ D_k) + \hat{\alpha}_{MH}(B_k\ C_k)$ and $V_k = (A_k + D_k) + \hat{\alpha}_{MH}(B_k + C_k)$.

The Mantel-Haenszel chi-square test of the null hypothesis of no DIF was not examined in this research.

4.2. Standardization Analysis

The standardization DIF measure, developed by Dorans and Kulick (1986), is

$$STD\ P-DIF = \hat{p}_F - \tilde{p}_R \tag{6}$$

where $\hat{p}_F$ is the proportion in the focal group who get the studied item correct, and $\tilde{p}_R$ is an adjusted proportion correct on the item for the reference group, defined as

$$\tilde{p}_R = \sum_k \frac{n_{Fk}}{n_F} \hat{p}_{Rk} \tag{7}$$

where $n_F = \sum_k n_{Fk}$ is the total number of examinees in the focal group and $\hat{p}_{Rk} = \frac{A_k}{n_{Rk}}$ is the proportion correct for the reference group in the k^{th} level of the matching variable.[1] One interpretation of $\tilde{p}_R$ is that it is the proportion of reference group examinees who would have got the studied item right if the distribution of the matching variable in the reference group had been the same as in the focal group.

The estimated standard error for *STD P – DIF* is given by the formula

$$SE(STD\,P-DIF) = \sqrt{\sigma_F^2 + \sigma_R^2} \tag{8}$$

where

$$\sigma_F^2 = \frac{1}{n_F} \hat{p}_F (1 - \hat{p}_F) \tag{9}$$

and

$$\sigma_R^2 = \frac{1}{n_F^2} \sum_k \frac{n_{Fk}^2}{n_{Rk}} \hat{p}_{Rk} (1 - \hat{p}_{Rk}) \tag{10}$$

4.3. Definition of Sample Size Conditions for DIF Analysis

The difficulty of defining sample sizes in a CAT simulation was addressed in ZTW. If groups of a fixed sample size had been drawn and the CAT administered, the sample sizes *per item* would have had a huge range. Because the goal was to investigate the behavior of DIF statistics

for specific sample sizes, it would not have been useful simply to analyze the available data for each item. After considering several other approaches, including resampling techniques and multiple replications, the *expected table* (ET) method was adopted:[2] First, within each simulation condition, item response data were generated for 60,000 examinees per group. Each examinee received 25 of the 75 CAT pool items. For each of the 75 items, all the available CAT data (out of a maximum of 60,000 responses per group) were then used to form the 2 (item responses) x 2 (groups) x *K* (levels of the matching variable) contingency table needed for DIF analysis. (The response frequency per item ranged from about 1,800 to about 34,000.) The table frequencies were then converted to proportions of the total number of observations for the group in question. Using these proportions as estimates of the population probabilities associated with the 4 x *K* cells for the relevant configuration of conditions, expected tables for the target sample sizes were obtained by multiplying the probability estimates for focal group cells by the desired focal group sample size and then doing the same for the reference group cells. Next, DIF statistics and standard errors were computed, based on the expected tables. A simple example of the ET approach, which originally appeared in ZTW, is given in the Appendix.

Though it produces only a single estimate, the ET approach can provide a relatively precise idea of the behavior of the *MH D-DIF* statistic. A supplementary study comparing the ET method to an estimation procedure based on multiple replications (as in a typical simulation study) appears in ZTW. The comparison was based on items for which 60,000 responses per population group were available. The ET method was found to give results similar to those of the replication-based approach. For the items that were studied, the ET-estimated *MH D-DIF* was determined to be as precise as an average over 316 replications of the *MH D-DIF* statistic based on the target sample sizes. Another advantage of the ET approach is that, once the 2 x 2 x *K* probability tables have been created, DIF results can be generated easily for any target sample size, facilitating further research.

The values of *MH D-DIF*, *SE*(*MH D-DIF*), *STD P-DIF*, and *SE(STD P-DIF)* in this report were computed from expected tables using equations 2 through 10. Note that *SE*(*MH D-DIF*) and

SE(STD P-DIF) are not indexes of the error associated with the *estimation* of *MH D-DIF* and *STD P-DIF*. Instead, these standard errors closely approximate the values of *SE(MH D-DIF)* and *SE(STD P-DIF)* that would be obtained using actual samples of the target sizes. The appropriate formulas for the standard errors of the ET estimates of *MH D-DIF* and *STD P-DIF* (which reflect the degree of precision with which the population DIF values are estimated using the ET approach) are given in ZTW. In the present study, item use patterns resulted in slightly smaller standard errors of the estimate than in ZTW. For *MH D-DIF*, these standard error values averaged 0.06 in both Pools, compared to 0.08 in ZTW. For *STD P-DIF*, they averaged 0.006, compared to 0.007 in ZTW.

5. Ability Estimation Results

First, the distributions of examinee residuals are described. Next, the effect of the observed estimation bias on ranking accuracy is discussed. DIF results appear in a subsequent section.

5.1. Distributions of Residuals for Rasch and 3PL Expected True Scores

Table 3 gives descriptive information on the distributions of residuals for four kinds of expected true scores. Two of the scores--3PL-CAT and Rasch-CAT--are based on adaptive administration. They were obtained according to equation 1, using the appropriate model-based estimates of item response functions and abilities. The other two scores, 3PL-75 and Rasch-75, are expected true scores evaluated at the ability estimate based on responses to all 75 items. These values represent the scores that would be obtained via nonadaptive administration. Again, estimates of the item response functions and examinee abilities were based on either the 3PL or Rasch model, as appropriate.

The residual for an examinee was obtained by subtracting the examinee's true score (based on the generating abilities and item parameters) from the value of the expected true score in question. The true score is defined as $\sum_{j=1}^{75} P_j(\theta)$, where $P_j(\cdot)$ denotes the 3PL function.

Insert Table 3 about here.

For each type of expected true score, Table 3 gives the median and the 25th and 75th percentiles of the residual distribution for the reference group and the three focal groups, based on samples of 2,000. In the case of the focal groups, results are given separately for each of the two pools. Within each combination of population and item pool, results are interdependent across the four types of residuals because the same sample of 2,000 examinees was used. The standard errors are approximately 0.11 for the medians and 0.06 for the quartiles. The standard deviations of the true scores ranged from 11.6 to 13 for the four populations; this provides an additional reference point for evaluating the size of the residuals.

The results in Table 3 are striking in their consistency: In each population-pool combination, the distribution of residuals for the first three scores looks acceptable, whereas the results for the Rasch CAT show a marked underestimation of examinee abilities. This is most striking in the Pool 1 results for the focal group with a mean of -1. Here, the examinees were underestimated by an average of one fifth of a standard deviation. A quarter of the examinees were underestimated by at least one half a standard deviation.

The results in Table 3 can be understood better through separate consideration of easy, medium, and difficult items. Figure 1 gives true and estimated test characteristic curves for all 75 items and the three panels of Figure 3 give (sub)test characteristic curves for easy (b = -1.95 or -1.30), medium (b = -0.65, 0, or .65), and difficult (b = 1.30 or 1.95) items. Each graph gives the true curve based on the generating parameters, the estimated curve based on the 3PL parameter estimates, and the estimated curve based on the Rasch parameter estimates. Figure 1 shows that the 3PL-estimated curve tracked the true curve well, while the Rasch curve was, as expected, quite discrepant for low ability values. Figure 2 shows the discrepancy to be severe for the hardest items and moderate for the middle-difficulty items. For the easy items, the Rasch curve was close to the true curve.

Insert Figures 1-2 about here.

Table 4 gives a detailed display of examinee residuals for the two types of Rasch-based expected true scores. Results are given for easy, medium, and difficult items for each quintile of the focal population that had a mean of -1. The results are in a different metric from those of Table 3: Because there were different numbers of easy, medium, and difficult items, the median residuals in the expected true score metric were divided by the number of items in the set to yield median "per-item" residuals. Results are shown for Pool 1; DIF, therefore, had no effect on these values. Standard errors of the tabled values range from approximately 0.001 to 0.004.

Insert Table 4 about here.

Some of the important aspects of the results in Table 4 are these:

- The degree to which underestimation occurred depended on the difficulty of the items and on the ability level of the examinees. Underestimation was worst for low-ability examinees and for difficult items.
- On the Rasch-75 test, overestimation on easy items tended to compensate for underestimation on hard items. On this nonadaptive test, the ability estimates are those that minimize the differences between observed percents correct and their model-based estimates *for all 75 items*. Because of model misspecification, the Rasch estimation procedure did not work well on individual items, but it did produce average residuals close to zero for expected true scores computed across the entire set of items. This was not true on the Rasch CAT, where the expected true scores tended to be underestimates, especially for examinees from the lowest quintiles. Because the CAT algorithm targets items to examinees, the ability estimates for these low-ability examinees are based primarily on responses to the easiest items. For the low-ability examinees, therefore, the estimation procedure minimizes the differences between observed and predicted item results *for the easy items*

only, and the residuals for these items were close to zero, as they should be. But when these ability estimates were used to project scores on the harder items, the Rasch model worked very poorly. For examinees with very low ability, the true probability correct is *c* for these items, whereas the Rasch-estimated proportion correct is close to zero. For these examinees, the average residual proportion correct, given in the lower-left corner of Table 4, was -.13 (-.12 for the nonadaptive test). This is consistent with the fact that the generating *c* value for the items was .15.

• For the lower three quintiles, CAT-based estimation was best on easy items and worst on difficult items, with medium items falling in between. In general, CAT estimation was best on items with difficulty levels that corresponded well with examinee abilities. For the two highest quintiles, however, CAT estimation worked quite well for all items: At higher ability levels, the test characteristic curve for the misspecified (Rasch) model was very close to the true test characteristic curve.

In summary, the Rasch model yielded acceptable results when responses on all the items were used to produce expected true scores on all the items. Results were also acceptable when results on a subset of the items (say, the easiest items) were used to produce expected true scores *on those items only.* Results were not acceptable when the CAT responses were used to estimate expected true scores for the entire pool.

The underestimation phenomenon illustrated in Tables 3 and 4 could be particularly problematic in a context in which Rasch CAT results were to be used not merely to rank examinees but to determine whether they exceeded a prescribed cut score. The downward bias that occurred in this simulation, however, may be less severe in CATs that select items on the basis of multiple criteria, rather than on the basis of item information alone. Constraints based on item content and item type, for example, may serve to moderate the estimation bias. A reduction in bias is to be expected because the ability range of the examinees receiving the more constrained CAT will be wider than the ability range for a CAT that is purely information-driven.

5.2 Effect of Rasch Estimation on Accuracy of Ranking Examinees

In some cases, the primary purpose of a test is to rank examinees. In this situation, it is not the minimization of bias, but the maximization of ranking accuracy that is of interest. Hambleton and Cook (1983) and Van de Vijver (1986) conducted simulation studies that included an examination of the effect on ranking accuracy of assuming the Rasch model when data have been generated using the 3PL model. (These studies concern nonadaptive tests; there have apparently been no published studies of model misspecification in the context of adaptive tests.) Hambleton and Cook (1983) found that, for examinees in the lower half of the simulated ability distributions, estimating ability with the Rasch model instead of the correct (3PL) model reduced the accuracy of examinee ranking to a considerable degree. For examinees in the upper half of the distributions, the Rasch model was about as effective as the 3PL for purposes of ranking examinees. Van de Vijver (1986) found that correlations between true and estimated abilities averaged about .75 when the Rasch model was fitted to 3PL data, compared to .89 when the Rasch model was fitted to Rasch data. (Records with all incorrect or all correct responses were deleted prior to model estimation.)[3]

The results of the present study appear quite consistent with those of Hambleton and Cook (1983). Table 5 gives some relevant results for the same sample that was analyzed to produce Table 4: 2,000 examinees from the focal group centered at -1, with Pool 1 items. The Spearman correlations of the four types of ability estimates with true ability are given for the entire sample and for the upper and lower halves of the distribution. Because the expected true score metric and the ability metric are monotonically related, these Spearman correlations are the same (except for possible differences in rounding error) as those that would be obtained using the expected true score metric.[4]

Insert Table 5 about here.

Results show that the Rasch-CAT had the worst ranking accuracy, although it is only slightly lower than the 3PL-CAT in the upper half of the ability distribution (.83, compared to .85). In the lower half, the difference is more substantial (.74, compared to .81). In the total group and in the upper half, the two 75-item tests yielded better accuracy than the two CATs. In the lower half, the 3PL CAT performed slightly better than the Rasch-75 test. One factor that must be taken into consideration when comparing the ranking accuracy of the nonadaptive tests to that of the CATs is reliability differences: The 75-item tests are based on three times as many item responses as the CATs. The results in Table 5 do not attempt to "correct" for this.

6. DIF Results

DIF analyses were conducted for the 12 simulation conditions of Table 1. In the first subsection, results are compared to those in ZTW and to the true DIF for the item. The second subsection, the classification of items into DIF categories is discussed.

6.1 Comparison of Mantel-Haenszel and Standardization Results to ZTW Results and to True DIF

For purposes of these analyses, the true DIF of an item was defined as the generating value of *ad* for that item (see Table 2), based on the following rationale: Under certain Rasch model conditions, the *MH D-DIF* statistic provides an estimate of 4*ad*. (See Donoghue, Holland and Thayer, 1993, who based their result on the work of Holland and Thayer, 1988.) The assumptions under which this finding holds are that (1) within each of the groups (reference and focal), the item response functions follow the Rasch model (obtained by setting $c_j = 0$ for all items j and $a_j \equiv a$ for all items j) , (2) the matching variable is the number-right score based on all items, *including* the item under analysis, referred to as the *studied item*, and (3) the items have the same item response functions for the reference and focal groups (i.e., $b_{jR} = b_{jF} \equiv b_j$), with the possible exception of the studied item. When conditions (1)-(3) do not hold, the population odds ratios will not, in general, be constant across number-right score levels (see Zwick, 1990). Therefore, for the 3PL model, it is not possible to derive a general expression for the quantity estimated by the *MH*

D-DIF statistic. However, empirical analyses showed that, in the ZTW study, *MH D-DIF* was approximately equal to 3*ad*.

For each of the item pools, Table 6 gives the means and standard deviations across the 75 items of *MH D-DIF*, *SE(MH D-DIF)*, *STD P-DIF*, and *SE(STD P-DIF)* for the two sample size conditions. An artifact of the simulation procedures needs to be considered when interpreting these results: DIF statistics for the two sample size conditions are based on the same ET probability tables. Because of this, the *MH D-DIF* measures are highly correlated across sample size conditions. The ET-estimated *STD P-DIF* statistic is invariant over target sample sizes. The primary purpose of including more than one sample size condition was to study the behavior of the standard errors of the DIF measures under various sample sizes.

Insert Table 6 about here.

In Pool 1, all values of *d* were zero; in Pool 2, the mean value of *ad* was 0.007 and the standard deviation was 0.293. Ideally, then, the mean of *MH D-DIF* and *STD P-DIF* should be equal to zero in both pools. The standard deviation of *MH D-DIF* should be close to zero in Pool 1 and roughly equal to 3(0.293) = 0.879 in Pool 2. Based on these criteria, results from the Rasch analysis appear quite acceptable; they are very similar to those obtained in ZTW using the 3PL model.[5]

The standard errors of the DIF statistics were nearly identical to those obtained in the ZTW study. For *MH D-DIF*, standard errors ranged from 0.3 to 0.4 for $n_R = 500$, $n_F = 500$ and from 0.5 to 0.7 for $n_R = 900$, $n_F = 100$. For *STD P-DIF* x 10, standard errors were approximately 0.3 for all $n_R = 500$, $n_F = 500$ conditions and about 0.5 for all $n_R = 900$, $n_F = 100$ conditions.

Table 7 gives the intercorrelations of the Rasch and 3PL DIF measures and true DIF for the $n_R = 500$, $n_F = 500$ sample size conditions in Pool 2. The results differed very little over focal group populations; the values shown are averages of the within-group correlations. None of these average correlations were below .97. Correlations were nearly identical for $n_R = 900$, $n_F = 100$.[6]

Insert Table 7 about here.

6.2 Expected Percent of "C" Results

ETS has a system for categorizing the severity of DIF based on MH results. According to this classification scheme, a "C" categorization, which represents extreme DIF, requires that the absolute value of *MH D-DIF* be at least 1.5 and be significantly greater than 1 (at $\alpha = .05$). A "B" categorization, which indicates moderate DIF, requires that *MH D-DIF* be significantly different from zero (at $\alpha = .05$) and that the absolute value of *MH D-DIF* be at least 1, but not large enough to satisfy the requirements for a C item. Items that do not meet the requirements for either the B or the C categories are labeled "A" items, which are considered to be free of DIF. Items that fall in the C category are typically eliminated from tests or subjected to further scrutiny.

Because most of the ET-estimated DIF statistics were based on at least 10,000 observations in the present study, it is reasonable to assume that they provide precise estimates of the population mean and standard deviation of *MH D-DIF* for the relevant configuration of item properties and simulation conditions. This is supported by analyses reported in ZTW. If it is assumed that *MH D-DIF* statistics for this configuration follow a normal distribution with this mean and standard deviation, percentiles of the theoretical distribution of *MH D-DIF* can be obtained. These percentiles can then be used to estimate the percent of times such an item would be classified as an A, B, or C if repeated random samples of observations were drawn. As detailed in ZTW, an algorithm was developed for estimating these percents, to be applied separately to each item in each condition. The algorithm was tested and found to work well with data from two simulation studies. This type of analysis allows an item's observed DIF status to be expressed in probabilistic, rather than deterministic terms, as an alternative way of providing information about the sampling variation of the *MH D-DIF* statistic.

In the 3PL model, the determination of which items are *nominally* A, B, and C items is not straightforward. Based on the empirical finding that *MH D-DIF* was approximately equal to 3*ad* in the conditions investigated in ZTW, items with *ad* = ±. 0.52 or ± 0.70 can be considered nominal C's; items with *ad* = ± 0.35 would be nominal B's; and items with *ad* = 0 or ±. 0.26 would be nominal A's. (See ZTW for further discussion of these issues.)

For Pool 2 conditions with n_R = 500, n_F = 500, Table 8 gives the average expected percent of C results for nominal A, B, and C items, respectively. Results are based on the 71 items that were administered in ZTW. (Four were never administered because there were always at least 25 more informative items.) Results obtained from the 3PL (ZTW) study and from the present study are given. The procedure for obtaining the values displayed in Table 8 was as follows: First, the expected percent of C results was estimated for each item in each simulation condition, using the DIF statistics obtained from the 3PL or Rasch study, as appropriate. Then, within each simulation condition, the results were averaged across items with the same nominal DIF status. Despite the high correlation (see Table 7) between the Rasch and 3PL DIF statistics, Table 8 shows that DIF was somewhat less likely to be detected in the Rasch analysis. The effect was largest for the focal group centered at -1, where the average detection rate for C items was 69% in the Rasch analysis, compared to 89% in the 3PL analysis. The discrepancy was smaller for the focal groups centered at 0 and 0.5. Examination of the DIF results showed a tendency for DIF statistics to be slightly smaller in magnitude in the Rasch analyses, particularly for the lowest-ability focal group. It is likely that the lower DIF detection rates for this focal group occurred because the quality of matching was particularly impaired when this group, with its substantial downward bias in CAT expected true scores, was paired with the reference group. It should also be noted, however, that the rate of falsely identifying Pool 2 items as C's was slightly higher in the 3PL than in the Rasch analysis. For Pool 1 items, the expected percent of C results was approximately zero in both the Rasch and the 3PL analyses.

Insert Table 8 about here.

7. Summary

In this simulation study of ability and DIF estimation in computer-adaptive tests, the Rasch model was fitted to data generated using the 3PL model. Rasch item parameter estimates determined the information table used in the item selection algorithm, and Rasch-based expected true scores on the CAT were obtained for each examinee. Examinees were matched on these expected true scores, and Mantel-Haenszel and standardization DIF analyses were then conducted.

Expected true scores from the Rasch-based CAT tended to be biased downward, particularly for lower-ability examinees. Not surprisingly, the nonadaptive and CAT expected true scores based on the correct (3PL) model behaved well. Rasch expected true scores based on nonadaptive administration of all pool items were also quite satisfactory. On the Rasch nonadaptive test, overestimation on easy items tended to compensate for underestimation on hard items. On the nonadaptive test, the ability estimates are those that minimize the differences between observed percents correct and their model-based estimates *for all 75 items*. Because of model misspecification, the Rasch estimation procedure did not work well on individual items, but it did produce average residuals close to zero for expected true scores computed across the entire set of items. On the CAT, however, ability estimates are based on responses to only a subset of items that have difficulties close to the examinee ability level. The estimation procedure minimizes the differences between observed and predicted item results *for these items only*, and the residuals for such items were typically close to zero, as desired. But when these ability estimates were used to project scores on other items, the misspecified (Rasch) model performed poorly.

The downward bias of Rasch CAT scores may be less serious in CATs that select items on the basis of multiple criteria, rather than on the basis of item information alone. Constraints based on item content and item type, for example, may serve to moderate the estimation bias that was

evident in this simulation. Testing programs using Rasch-based CATs may wish to conduct simulations, using their own CAT algorithms, to further investigate the bias phenomenon. For programs that use cut scores, it may be useful to focus on misclassification rates, rather than on the expected magnitude of the bias.

The Rasch-based DIF statistics were found to be highly correlated with the generating DIF, and the means and standard deviations of these statistics across items were close to their nominal values. There was, however, a tendency for DIF statistics to be slightly smaller in magnitude than in the 3PL analysis, resulting in a lower probability of detecting items with extreme DIF, particularly when the reference and focal groups had different ability distributions. It is likely that this reduced sensitivity resulted from the degradation in the accuracy of matching that occurs when the ability estimates for the two groups are biased to different degrees.

References

Birnbaum, A. (1968). Some latent trait models. In Lord, F. & Novick, M. (1968), *Statistical theories of mental test scores*, pp. 397-424. Reading, MA: Addison-Wesley.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland and H. Wainer (eds.) *Differential Item Functioning*, pp. 137-166. Hillsdale, NJ: Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.

Hambleton, R. K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In Weiss, D. J. (Ed.), *New Horizons in Testing: Latent Trait Theory and Computerized Adaptive Testing*. New York: Academic Press.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty*. ETS Research Report No. RR 85-43. Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test Validity*, pp. 129-145. Hillsdale, NJ: Erlbaum.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

McHale, F., Dorans, N., Holland, P., & Petersen, N. (May 2, 1988). Specifications for standardized percent correct and distractor analysis (IANA80 and IANA82). Technical memorandum, Educational Testing Service.

Phillips, A. & Holland, P. W. (1987). Estimation of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics, 43*, 425-431.

Robins, J., Breslow, N., & Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics, 42*, 311-323.

Van de Vijver, F. J. R. (1986). The robustness of Rasch estimates. *Applied Psychological Measurement, 10*, 45-57.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (ed.), *Applications of item response theory*. Vancouver: Educational Research Institute of British Columbia.

Wingersky, M. S., Patrick, R., & Lord, F. M. (1988). *LOGIST user's guide: LOGIST Version 6.00*. Princeton, NJ: Educational Testing Service.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185-197.

Zwick, R., Thayer, D. T., & Wingersky, M. (in press). A simulation study of methods for assessing differential item functioning in computer-adaptive tests. *Applied Psychological Measurement.*

Zwick, R., Thayer, D. T., & Wingersky, M. (1993). *A simulation study of methods for assessing differential item functioning in computer-adaptive tests.* ETS Research Report 93-11. Princeton, NJ: ETS.

Footnotes

[1]When n_{Rk} is equal to zero, both $\tilde{p}_R$ and σ_R^2 are undefined. When this occurs, the standard ETS DIF software implements an imputation procedure proposed by Holland, (McHale, Dorans, Holland & Petersen, 1988). A modification of this procedure which takes into account the special nature of the CAT-based analyses was used in this study.

[2]This approach was proposed by Charles Lewis.

[3]Van de Vijver also reported on the bias of the ability estimates. Because of the arbitrary scaling of the estimates, however, these results do not appear to be meaningful.

[4]For some examinees for whom estimation problems occurred, ability estimates of -5 were assigned. For the sample displayed in Table 5, this occurred for 1.1% of the 3PL-75 estimates and 1.7% of the 3PL-CAT estimates. No ability estimates of -5 occurred for the Rasch-75 or the Rasch-CAT estimates.

[5]When DIF analyses are conducted in which the matching variable is the number-right score, including the studied item, the *STD P-DIF* statistics are constrained to sum to zero across items; the *MH D-DIF* statistics are also constrained to sum to approximately zero. This constraint does not apply in the present analyses; that is, the closeness to zero of the means of Table 6 was not a foregone conclusion.

[6]Reliabilities were computed for each DIF statistic using the methodology described in ZTW. Because they were all at least .997, reliability-corrected results were essentially identical to the displayed (uncorrected) results.

Appendix: Example of the Expected Table (ET) Approach to Estimation of DIF Statistics

Consider the following hypothetical results for a single item, assuming that there are only two levels of the matching variable. The first step is to use all the data available for the item to construct a 2 x 2 x 2 frequency table (because $K = 2$ here). Then the cell frequencies for the reference group are divided by the total number of reference group examinees and the frequencies for the focal group are divided by the total number of focal group examinees, producing the following 2 x 2 x 2 table of probabilities:

Low on Matching Variable

	Right	Wrong	Total
Reference	.2	.1	.3
Focal	.2	.2	.4

High on Matching Variable

	Right	Wrong	Total
Reference	.5	.2	.7
Focal	.4	.2	.6

Now suppose that target tables are needed for the $n_R = 900$, $n_F = 100$ condition. The reference group probabilities are multiplied by 900 and the focal group probabilities are multiplied by 100, producing the following table for use in DIF analysis.

Low on Matching Variable

	Right	Wrong	Total
Reference	180	90	270
Focal	20	20	40

High on Matching Variable

	Right	Wrong	Total
Reference	450	180	630
Focal	40	20	60

Table 1

The 12 Administration Conditions

		Sample size per item		
Condition	Focal Group[a]	Focal	Reference	Pool
1	N(-1,1)	100	900	1
2	N(-1,1)	500	500	1
3	N(-1,1)	100	900	2
4	N(-1,1)	500	500	2
7	N(0,1)	100	900	1
8	N(0,1)	500	500	1
9	N(0,1)	100	900	2
10	N(0,1)	500	500	2
13	N(.5,1)	100	900	1
14	N(.5,1)	500	500	1
15	N(.5,1)	100	900	2
16	N(.5,1)	500	500	2

Note. These 12 conditions are a subset of those used in ZTW. The original numbering has been retained.

[a]"N(x,y)" indicates a normal distribution with mean x and standard deviation y.

Table 2

True *a*, *b*, and *d* Parameters and Rasch Estimates of *b* Parameters

Item	Pools 1 and 2		Pool 2		
	a	*b*	*d*	*ad*	$\hat{b}$
1	0.74	-1.95	-0.35	-0.26	-2.37
2	0.74	-1.95	-0.35	-0.26	-2.37
3	0.74	-1.95	0.00	0.00	-2.24
4	0.74	-1.95	0.00	0.00	-2.30
5	0.74	-1.95	0.35	0.26	-2.37
6	0.74	-1.30	-0.35	-0.26	-1.71
7	0.74	-1.30	-0.35	-0.26	-1.77
8	0.74	-1.30	0.00	0.00	-1.61
9	0.74	-1.30	0.00	0.00	-1.64
10	0.74	-1.30	0.35	0.26	-1.59
11	0.74	-1.30	0.35	0.26	-1.71
12	0.74	-0.65	-0.35	-0.26	-0.99
13	0.74	-0.65	-0.35	-0.26	-0.94
14	0.74	-0.65	0.00	0.00	-1.12
15	0.74	-0.65	0.00	0.00	-0.99
16	0.74	-0.65	0.00	0.00	-1.03
17	0.74	-0.65	0.35	0.26	-0.97
18	0.74	-0.65	0.35	0.26	-1.03
19	0.74	0.00	-0.70	-0.52	-0.34
20	0.74	0.00	-0.35	-0.26	-0.39
21	0.74	0.00	0.00	0.00	-0.33
22	0.74	0.00	0.00	0.00	-0.35
23	0.74	0.00	0.35	0.26	-0.36
24	0.74	0.00	0.35	0.26	-0.33
25	0.74	0.00	0.70	0.52	-0.38

(continued)

Table 2 (continued)

True a, b, and d Parameters and Rasch Estimates of b Parameters

Item	Pools 1 and 2		Pool 2		
	a	b	d	ad	$\hat{b}$
26	0.74	0.65	-0.35	-0.26	0.27
27	0.74	0.65	0.00	0.00	0.25
28	0.74	0.65	0.00	0.00	0.21
29	0.74	0.65	0.35	0.26	0.22
30	0.74	0.65	0.35	0.26	0.28
31	0.74	0.65	0.70	0.52	0.28
32	0.74	1.30	-0.70	-0.52	0.74
33	0.74	1.30	-0.35	-0.26	0.71
34	0.74	1.30	0.00	0.00	0.87
35	0.74	1.30	0.00	0.00	0.71
36	0.74	1.30	0.35	0.26	0.74
37	0.74	1.95	0.00	0.00	1.21
38	0.74	1.95	0.35	0.26	1.22
39	1.00	-1.95	-0.35	-0.35	-2.68
40	1.00	-1.95	0.00	0.00	-2.73
41	1.00	-1.30	-0.35	-0.35	-1.96
42	1.00	-1.30	0.00	0.00	-2.02
43	1.00	-1.30	0.00	0.00	-1.89
44	1.00	-1.30	0.35	0.35	-1.96
45	1.00	-0.65	-0.35	-0.35	-1.06
46	1.00	-0.65	-0.35	-0.35	-1.12
47	1.00	-0.65	0.00	0.00	-1.17
48	1.00	-0.65	0.00	0.00	-1.10
49	1.00	-0.65	0.35	0.35	-1.09
50	1.00	-0.65	0.70	0.70	-1.21

(continued)

Table 2 (continued)

True a, b, and d Parameters and Rasch Estimates of b Parameters

Item	Pools 1 and 2		Pool 2		
	a	b	d	ad	$\hat{b}$
51	1.00	0.00	-0.70	-0.70	-0.40
52	1.00	0.00	-0.35	-0.35	-0.39
53	1.00	0.00	-0.35	-0.35	-0.36
54	1.00	0.00	0.00	0.00	-0.31
55	1.00	0.00	0.00	0.00	-0.36
56	1.00	0.00	0.35	0.35	-0.31
57	1.00	0.00	0.70	0.70	-0.41
58	1.00	0.65	-0.35	-0.35	0.28
59	1.00	0.65	-0.35	-0.35	0.40
60	1.00	0.65	0.00	0.00	0.37
61	1.00	0.65	0.00	0.00	0.35
62	1.00	0.65	0.00	0.00	0.33
63	1.00	0.65	0.35	0.35	0.32
64	1.00	0.65	0.35	0.35	0.26
65	1.00	1.30	-0.35	-0.35	0.95
66	1.00	1.30	-0.35	-0.35	0.90
67	1.00	1.30	0.00	0.00	0.91
68	1.00	1.30	0.00	0.00	0.93
69	1.00	1.30	0.35	0.35	0.88
70	1.00	1.30	0.35	0.35	0.95
71	1.00	1.95	-0.35	-0.35	1.48
72	1.00	1.95	0.00	0.00	1.32
73	1.00	1.95	0.00	0.00	1.33
74	1.00	1.95	0.35	0.35	1.39
75	1.00	1.95	0.35	0.35	1.38

Note. The true c parameters were .15 for all items for both pools. The true d parameters were zero in Pool 1. The estimated b parameters were obtained using a reference group (Pool 1) sample of 2,000.

Table 3

Percentiles of Distributions of Residuals
for Four Types of Expected True Scores
(Samples of 2,000)

	Pool 1 (No DIF)			Pool 2		
	25th %tile	Median	75th %tile	25th %tile	Median	75th %tile
Reference Group						
3PL - 75[a]	-2.4	-0.1	2.4			
3PL - CAT	-2.8	-0.1	2.9			
Rasch - 75	-2.4	0.0	2.4			
Rasch - CAT	-4.8	-1.3	2.0			
Focal N(-1,1)						
3PL - 75	-2.2	-0.0	2.1	-2.2	0.0	2.1
3PL - CAT	-2.9	-0.3	2.5	-2.8	-0.3	2.7
Rasch - 75	-2.4	-0.1	2.3	-2.3	0.2	2.7
Rasch - CAT	-6.1	-2.6	1.5	-6.2	-2.2	1.9
Focal N(0,1)						
3PL - 75	-2.1	0.0	2.2	-2.0	0.2	2.3
3PL - CAT	-2.9	0.0	3.0	-2.8	-0.0	3.0
Rasch - 75	-2.1	0.1	2.3	-2.0	0.3	2.7
Rasch - CAT	-5.0	-1.4	2.2	-4.4	-0.8	2.9
Focal N(.5,1)						
3PL - 75	-2.3	-0.1	2.1	-2.0	0.0	2.0
3PL - CAT	-2.9	-0.1	2.6	-2.7	-0.1	2.8
Rasch - 75	-2.2	-0.0	2.3	-1.9	0.2	2.2
Rasch - CAT	-4.3	-1.1	1.9	-3.9	-0.6	2.5

Note. Residuals were computed by subtracting the true score (computed using the generating abilities and item parameters) from the expected true score. The standard deviation of the true scores ranged from 11.6 to 13.0 for the four population groups. Standard errors of the tabled values are approximately 0.11 for medians and 0.06 for the 25th and 75th percentiles.

[a] "75" denotes the expected true score obtained via nonadaptive testing.

Table 4

Median Residual per Item for Subsets of Items and Examinees:
Focal N(-1,1) Group, Pool 1 (N=2,000)

			Quintile of Ability Distribution				
Type of Item	Number of Items		1	2	3	4	5
Easy	17	Rasch-75	.16	.07	.03	-.00	-.01
		Rasch-CAT	.02	.01	-.01	-.01	-.02
Medium	40	Rasch-75	-.02	.00	.02	.02	-.00
		Rasch-CAT	-.08	-.05	-.02	.00	-.02
Difficult	18	Rasch-75	-.12	-.10	-.07	-.03	.03
		Rasch-CAT	-.13	-.11	-.09	-.05	.01

Note. Each table entry is the median value for that group of examinees of $\frac{1}{n}\left[\Sigma \hat{P}(\hat{\theta}_{CAT}) - \Sigma P(\theta)\right]$, where the summation is over the items in that set (Easy, Medium, Difficult) and n is the number of items in the set. Standard errors of the tabled values range from approximately 0.001 to 0.004. Quintiles are ordered from lowest ability (1) to highest ability (5).

Table 5

Spearman Correlations of Estimated Abilities with True Ability:
Focal N(-1,1) Group, Pool 1 (N = 2,000)

	Total Group	Upper Half	Lower Half
3PL - 75	.96	.90	.84
3PL - CAT	.94	.85	.81
Rasch - 75	.95	.90	.79
Rasch - CAT	.91	.83	.74

Table 6

Means and Standard Deviations of DIF Statistics

	Pool 1						Pool 2					
Condition	1	2	7	8	13	14	3	4	9	10	15	16
Focal Group	N(-1,1)		N(0,1)		N(.5,1)		(N-1,1)		N(0,1)		N(.5,1)	
MH D-DIF												
mean	-.01	-.01	.00	.00	.01	.01	.00	.00	.02	.01	.01	.01
s.d.	.08	.08	.07	.07	.08	.08	.90	.92	.96	.96	1.0	.99
SE(MH D-DIF)												
mean	.54	.33	.54	.32	.55	.33	.55	.33	.55	.33	.56	.33
s.d.	.13	.08	.14	.08	.14	.09	.14	.08	.14	.08	.15	.09
STD P-DIF x 10												
mean	-.01	-.01	.00	.00	.01	.01	.01	.01	.01	.01	.00	.00
s.d.	.07	.07	.05	.05	.06	.06	.74	.73	.77	.77	.77	.77
SE(STD P-DIF) x 10												
mean	.49	.30	.48	.28	.47	.27	.49	.30	.47	.27	.46	.27
s.d.	.12	.07	.12	.07	.12	.07	.12	.07	.12	.07	.12	.07

Note. Pool 1 has no DIF. In Pool 2, the generating DIF has a mean of 0.007 and a standard deviation of 0.293 across items. Odd-numbered conditions have $n_R = 900$, $n_F = 100$; even-numbered conditions have $n_R = 500$, $n_F = 500$.

Table 7

Correlations of Rasch and 3PL DIF Measures and True DIF:
Pool 2, $n_R = 500$, $n_F = 500$

		1	2	3	4	5
1	*MH D-DIF* - 3PL	1	.98	.99	.97	.98
2	*MH D-DIF* - Rasch		1	.98	1.00	.99
3	*STD P-DIF* - 3PL			1	.97	.98
4	*STD P-DIF* - Rasch				1	.98
5	True DIF (*ad*)					1

Table 8

Average Expected Percent of "C" Results: Pool 2, $n_R = 500$, $n_F = 500$

Condition		4		10		16	
Focal Group		N(-1,1)		N(0,1)		N(.5,1)	
Model		3PL	Rasch	3PL	Rasch	3PL	Rasch
Nominal Category	Number of Items						
A	45	1.3	0.9	1.3	1.4	1.4	1.6
B	20	8.2	7.5	11.8	9.4	14.2	10.5
C	6	89.2	89.3	87.2	78.5	88.8	82.5

Note. An example of interpretation, based on Condition 4, is as follows: On the average, the 6 nominal C items were expected to be correctly identified 89.2% of the time using the 3PL-based procedures and 69.3% of the time using the Rasch-based procedures.

Figure 1
Test Characteristic Curves

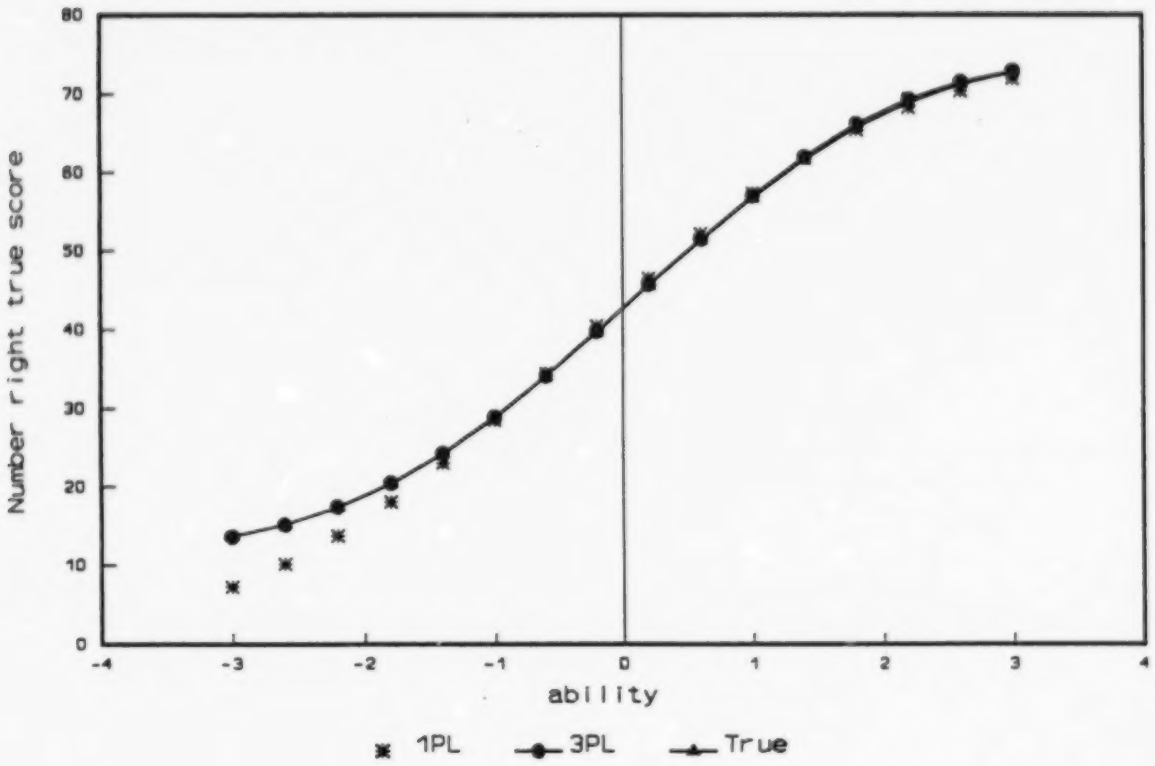

Figure 2
Test Characteristic Curve